경영정보학연구
제16권 제2호
2006년 6월

사례기반 추론을 이용한 암 환자 진료비 예측 모형의 개발*

정 석 훈**, 서 용 무***

Development of a Medial Care Cost Prediction Model for Cancer Patients Using Case-Based Reasoning

Sukhoon Chung, Yongmoo Suh

Importance of Today's diffusion of integrated hospital information systems is that various and huge amount of data is being accumulated in their database systems. Many researchers have studied utilizing such hospital data. While most researches were conducted mainly for medical diagnosis, there have been insufficient studies to develop medical care cost prediction model, especially using machine learning techniques.

In this research, therefore, we built a medical care cost prediction model for cancer patients using CBR (Case-Based Reasoning), one of the machine learning techniques. Its performance was compared with those of Neural Networks and Decision Tree models. As a result of the experiment, the CBR prediction model was shown to be the best in general with respect to error rate and linearity between real values and predicted values. It is believed that the medical care cost prediction model can be utilized for the effective management of limited resources in hospitals.

Keywords : Medical Data Mining, Cost Prediction Model, Case-Based Reasoning

* 본 연구는 2004년 고려대학교 연구비 지원 사업에 의한 것임. 적절한 지적으로 본 논문의 완성도를 높여주신 두 익명의 심사위원들에게 심심한 감사를 표합니다.
** 고려대학교 경영대학 박사과정
*** 고려대학교 경영대학 교수

Ⅰ. 서 론

정보시스템의 이용은 비단 일반 기업들에서뿐만 아니라, 각급 의·병원들에서도 점차 일반화되고 있는 추세이다. 초기에 수납 시스템 중심으로 이용되던 모습에서 한 걸음 더 나아가, 최근에는 진료 시스템 및 각종 첨단 의학 장비들과의 연결을 통해 통합된 병원 정보시스템을 구축하여 활용하고 있다. 대형 병원을 중심으로 이루어지고 있는 이러한 통합 병원 정보시스템의 운영은 병원에서 발생하는 모든 종류 및 모든 형태의 데이터들이 데이터베이스 시스템에 통합 운영되고 있다는 점에서 그 의미가 매우 크다고 할 수 있다.

병원 정보시스템의 데이터베이스에는 환자들의 증상에 관련된 데이터뿐만 아니라 각 환자들에게 시술 된 치료법 및 처방에 대한 데이터, 그리고 진료비 구성에 관련된 데이터 등 다양한 종류의 데이터들이 담기게 된다. 뿐만 아니라 병원에서의 정보시스템 이용은 각급 병원들로 하여금 대용량의 데이터를 확보할 수 있게 해주는 계기가 되었다. 따라서 많은 학자들이 이렇게 확보된 대용량의 데이터에 데이터마이닝 기법을 적용하여 다양한 분석을 시도하였으며, 그 동안 주목할 만한 결과들을 얻어내었다.

Mitchell 등[1999]은 데이터마이닝의 다양한 사용 분야에 대해서 언급하면서 의료 분야를 중요한 분야 중 하나로 지적하였고, Demšar 등[2001]은 속성 선택과 기계 학습 기법을 이용하여 환자들의 최종 상태를 예측 하는 모형을 개발함으로써 소수의 속성들이 모형에서 중요한 역할을 하고 있는 것을 발견해 내었다. Masuda 등[2002]은 데이터마이닝 기법을 이용하여 대용량의 의료 데이터베이스에서 자동으로 데이터를 분석하여 일정한 패턴을 발견해 주는 프레임워크를 제시 했으며, Jerez-Aragonés 등[2003]은 인공신경망과 의사결정나무를 이용하여 매우 유용하게 쓰일 수 있는 유방암 진단 모형을 개발 하였다.

이러한 연구들의 결과는 의학적인 연구나 병원의 경영에 이용되고 있다[Jenn-Lung 등, 2001]. 대표적으로 가장 많이 이용된 분야는 질병의 진단 및 예측 등에 관련된 분야로서 지금까지의 거의 대부분의 의료 데이터마이닝에 관련된 연구는 이 부분에 속한다. 이러한 연구는 의사들로 하여금 진단과 처방을 좀 더 정확하고 빠르게 수행할 수 있도록 도움을 줄 수 있다. 이에 비해서, 병원의 운영 및 관리를 위한 측면에서의 데이터마이닝 기법 이용은 상대적으로 적은 수의 연구가 수행되었으며 특히, 진료비 예측에 관련된 문제에 대해서는 그 수가 더욱 적다. 진료비에 대한 정확한 예측은 병원의 한정된 자원을 효율적으로 사용할 수 있게끔 도와주는 등, 병원의 각종 정책입안에 주요하게 사용될 수 있으며, 환자가 진료비에 대한 재정적인 계획을 세우는 데도 큰 도움을 줄 수 있게 한다[Fireman 등, 2000; Roche 등, 2002].

지금까지의 진료비 예측 모형들은 대부분 회귀분석이나 분산분석과 같은 통계적 기법을 이용한 연구들이 대부분이었다. 그러나 이러한 통계적 기법을 사용한 연구들은 기법의 특성상 입력 변수로 사용될 수 있는 데이터의 형태가 제한되어 있거나, 또는 너무 적은 수의 입력 변수만을 이용할 수 있기 때문에 좀 더 정확한 예측을 하는데 많은 제약이 따랐다[Tollestrup 등, 2001; Penberthy 등, 1999].

따라서 본 연구에서는 기법의 특성상 입력 변수와 출력 변수의 형태와 수에 비교적 덜 제한적인 방법인 사례기반추론(CBR: Case-Based Reasoning)을 이용하여 우리나라 4대 암(위암, 폐암, 간암, 그리고 대장암) 환자의 진료비 예측모형을 개발하고자 한다. 각 암 별로 예측모형을 개발할 것이며, 사례기반 추론을 이용하여 개발된 예측모형의 성능을 인공신경망 모형 및 의사결정나무 모형의 성능과 비교할 것이다. 여러 가지 암 중에서도 특히 4대 암에 대한 진료비를 예측 대상으로

삼은 이유는 위암, 폐암, 간암, 그리고 대장암이 우리나라의 사망 원인 질병 중 가장 높은 비중을 차지하고 있기 때문이다[한국 중앙 암 등록 본부, 2003]. 따라서 그만큼 예측 모형의 필요성 및 효용성이 가장 높다고 할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 이어지는 다음 장에서는 진료비 예측 모형에 대한 문헌들을 고찰하였다. Ⅲ장에서는 본 연구에서 사용된 데이터마이닝 기법들을 살펴보았고, Ⅳ장에서는 본 연구에서 사용된 데이터에 대한 설명, 그리고 실험에 사용된 예측 모형에 대해서 구체적으로 기술하였다. Ⅴ장에서는 개발된 예측 모형의 실험 결과 및 해석에 대해서 살펴 본 후, Ⅵ장에서는 결론 및 향후 연구 과제를 제시하였다.

Ⅱ. 진료비 예측 모형

회귀분석이나 분산분석과 같은 통계적 기법을 이용하거나 또는 인공 지능 기법들을 이용한 진료비 관련 분석 모형들이 지금까지 소수나마 시도되어 왔다. Ismael 등[1998]은 인공신경망을 이용하여 관상동맥 증후군 환자들을 대상으로 그 치료 비용을 예측하는 모형을 개발하였다. 총 4개의 모형을 생성하여 실험에 사용하였는데, 첫 번째 모형은 목표 변수인 비용을 0에서 1사이의 값으로 변환시켜주는 정규화 방법으로 선형 변환 방법을 사용한 것이며, 두 번째 모형은 비용에 로그변환을 시켜 준 것이다. 세 번째 모형에서는 훈련용 데이터의 비용을 이용하여 데이터를 세 개의 그룹(낮은 그룹, 중간 그룹, 그리고 높은 그룹)으로 나누고 각 그룹에 각각 인공신경망 모델을 따로 생성하였다. 따라서 검증용 데이터가 입력되면 일단 어떤 그룹에 속하는지에 대해서 판별한 후 다시 해당 그룹의 인공신경망 모형으로 비용을 예측하는 구조로 되어 있다. 비용 값의 정규화는 첫 번째 모형과 같은 선형 변형 방법을 사용하였다. 네 번째 모형은 세 번째 모형과 동일하지만 목표 변수인 비용 값의 정규화 방법을 로그 변환으로 사용한 모형이다. 입력 변수로는 환자의 상태, 그리고 합병증 등과 관련된 변수로 총 16개가 사용되었으며 이 중에서 비용을 로그함수로 변환 시켜 준 모형이 비교적 좋은 성능을 보이는 것으로 나타났다.

Penberthy 등[1999]은 유방암, 대장암, 폐암, 그리고 전립선 암 환자들 중 주로 노인들을 대상으로 그 진료비를 예측하는 회귀 모형을 개발하였다. 이 모형에는 총 6개의 입력 변수가 사용되었으며, R^2값은 0.38~0.49에 머물렀다.[1] 분산 분석을 사용한 연구로는 Tollestrup 등[2001]이 수행한 연구가 있다. 이 들은 유방암 비용에 있어서 히스페닉계 사람들과 비히스페닉계 사람들 간에 유의한 차이가 있는지를 밝히는 데 분산분석을 사용하였다. Goldman 등[2001]은 암환자들을 대상으로 임상적 치료를 받은 환자들과 그렇지 않은 환자들 간에 그 치료비에 있어 유의한 차이가 있는지를 밝히는 연구를 통계적 기법을 이용하여 수행하였다.

진료비를 직접 예측하는 모델 이외에 특정 요인과 진료비와의 관계를 밝히는 연구들도 수행되어 왔다. 그 중에서 몇 가지를 살펴 보면, 먼저 DeBerard 등[2003]은 미국 Utah 주의 직장 의료보험 환자들을 대상으로 생리심리사회학적 모델(biopsychosocial model)의 변수들과, 요추의 외과적 수술 비용 및 의료 보험 비용과의 관련성에 대한 연구를 수행하였다. 생리심리사회학적 모델은 총 12개의 변수로 구성되어 있는데, 이들은 생물학적 변수, 심리학적 변수, 그리고 사회학적 변수로 나뉘어 진다. 이들 변수들과 총 의료 비용과의 관계는 피어슨 상관계수로 그 관계를 살펴 보았으며, 의료 보험 비용과 총 의료 비용과의 관계는 회귀 모형으로 설명하고자 하였다. 그 결과 피어슨 상관계수의 경우 -0.375에서 0.314로 나타났는데 이는 두 변수간의 선형적 관계가 명확히 존재한다고 보기에는 어려운 수치이다. 회귀모형의

1) 이는 6개의 입력 변수가 진료비의 39~49% 밖에 못 미치는 부분을 설명해 주고 있다는 것이다.

R^2값은 0.37로 나타났는데 이는 의료 보험 비용이 총 의료 비용의 37%를 설명해 줄 수 있다는 의미이므로, 총 의료 비용은 의료 보험 비용 이외의 다른 요인들에 의해서도 많은 영향을 받고 있음을 알 수 있었다.

Sheng 등[2005]은 '병원 내 감염'의 문제가 의료기관의 형태에 따라 치료비, 입원기간, 그리고 치료 결과에 대해 어떤 영향을 주는지 분석하였다. 비교 되었던 두 형태의 의료기관은 대형 메디컬 센터와 중형 규모의 지역 의료기관이었다. 비교 연구를 수행한 결과 병원 내 감염에 대하여 질병의 심각성, 병원 내 감염 사례의 다양성 등은 대형 메디컬 센터에서 문제가 더 심각한 것으로 분석되었으나 입원 기간의 연장, 추가 비용의 증가 등은 의료 기간의 형태에 따라서 크게 다르지 않은 것으로 나타났다.

Jayadevappa 등[2005]은 전립선 환자들을 대상으로, 종족에 따른 다양한 환자 치료비의 변화에 대해서 분석하였다. 60명의 백인과 60명의 흑인으로 구성된 총 120명의 전립선 암 환자들과, 이들과 나이와 종족을 유사하게 구성한 240명의 대조군을 분석에 사용하였다. 종족간 인구통계학적 변수, 의학적 변수 그리고 치료 패턴들은 t-test와 χ^2분석을 이용하여 비교하였으며, 전립선 암 비용의 변화를 분석하고 종족과 의료비의 연관관계는 회귀모형을 사용하여 분석하였다. 분석 결과, 백인들이 흑인들 보다 좀 더 근본적인 전립선 절제술을 많이 시술 받는 것으로 나타났으며, 비용에 있어서는 대조군보다 실험군이 약 1.3배 높은 상승 비용을 보이는 것을 발견할 수 있었다. 또한 CCI(Charlson comorbidity Index)[2]가 치료와 비용에 대한 좋은 예측치로 사용될 수 있음을 증명하였다. 그리고 총 의료비는 환자의 종족과는 무관한 것으로 나타났다.

위에서 볼 수 있는 것과 같이, 진료비 예측에 관련된 연구는 아직 충분히 수행되지 않고 있으며, 지금까지 이루어진 연구들 대부분의 경우도 통계적 기법을 이용한 연구가 수행되었거나 진료비와 다른 요인들과의 관계를 밝히는 모형들이 개발된 정도이다.

Ⅲ. 본 연구에 사용된 데이터마이닝 기법 소개

3.1 인공신경망(Artificial Neural Networks)

인공신경망 기법은 가장 많이 사용되고 있는 기계학습 방법 중의 하나로, 인간의 무수히 많은 뇌세포들이 신경망을 통하여 주변의 많은 다른 뇌세포들과 정보를 주고 받으면서 계산을 수행하거나 의사결정을 내리는 과정을 입력 층, 은닉 층 그리고 출력 층으로 이루어진 간단한 구조로 단순화하여 모방하고 있는 기법이다. 따라서 이 기법을 사용하기 위해서는 먼저 뇌의 구조를 모방한 인공신경망의 구조를 정의하여야 하고, 뇌에서의 학습과정과 같이 인공신경망에서도 학습과정을 거쳐야 예측이나 분류를 할 수 있는 모델이 만들어지게 된다.

먼저, 인공신경망의 구조는 입력 층의 노드의 수, 은닉 층의 수와 각 은닉 층의 노드의 수, 그리고 출력 층의 노드의 수에 의하여 정의된다. 다음, 인공신경망에서의 학습은 인접한 다른 층의 노드와 노드를 이어주는 아크(arc)들의 가중치를 적절한 값으로 조정하는 과정을 통하여 이루어진다. 이 학습과정에서의 중요한 점은 가중치들을 조절할 때 어느 정도씩 어느 방향으로 조절할 것인가를 결정하는 것과, 또 언제 이 학습과정을 멈추도록 할 것인가를 지정하는 일이다.

상용 툴에서는 여러 가지 옵션을 제공함으로써 사용자가 원하는 인공신경망의 구조를 결정할 수 있도록 하고, 학습과 관련된 내부 변수들의 값을 변경해 봄으로써 학습과정을 조절할 수 있도록 하며 학습을 멈추기 위한 기준도 사용자가 지정하도록 하고 있다.

2) CCI: 다양한 의학적 만성 상태를 이용하여 사망률을 산정하기 위해서 사용되는 index.

인공신경망 기법은 noisy data에 크게 영향 받지 않는 장점이 있으나, 의사결정나무에 비하여 설명력이 부족하다는 단점을 갖고 있다. 따라서 설명력을 보완하기 위해서 신경망으로부터 규칙을 도출하는 방법에 대한 연구도 있었다. 인공신경망을 이용한 연구 영역은 매우 넓은데 예를 들어, 서로 다른 데이터베이스에서 일치하는 속성들을 찾는 문제[Li 등, 2000], 주식 시장의 수익 예측 문제[Enke 등, 2005], 수력 발전소의 수온 예측 문제[Romero 등, 2005], 생물학에의 이용[Jeng 등, 2006], 그리고 의료적 의사결정 문제[Mangalampalli 등, 2006] 등에 이르기까지 예측과 분류 문제에서 매우 광범위하게 사용되고 있다.

3.2 의사결정나무(Decision Tree)

일반적으로 분류 작업을 위하여 많이 사용 되어온 의사결정나무 기법은 이질적인 사례들의 큰 집합을 점차 덜 이질적인 사례들의 여러 작은 집합으로 분할하는 과정을 반복적으로 수행함으로써 의사결정에 사용될 나무구조의 모형을 형성한다. 따라서 이 과정에서 고려해야 할 것은 나무구조를 형성하는 각 노드에서의 분할 기준으로 사용되는 변수와 그 값의 선택은 어떻게 할 것인가, 분할 과정은 언제 멈추게 할 것인가, 작성된 모델의 분류 정확도가 떨어지면 어떻게 해결할 것인가 등이다.

먼저, 분할 기준은 큰 집합(사례집합)이 여러 작은 집합으로 분할된 후 이들의 확률적 순수도의 합이 분할 전 집합의 순수도 보다 가장 크게 하는 변수와 그 값을 찾게 되는데 이때 지니 인덱스 또는 엔트로피 지수 등을 활용한다. 이러한 분할 작업을 반복적으로 수행하다가 분할된 집합에 속하는 사례들의 수가 사용자가 지정한 수치보다 작거나, 그 집합의 순수도가 사용자가 지정한 수치보다 크게 되면 분할 작업을 멈추게 된다. 이렇게 형성된 의사결정나무 모델을 테스트 데이터들을 가지고 검사하는 과정에서 분류 오류가 높은 노드들은 가지치기 과정을 통해 제거하는 작업을 수행하게 된다.

상용 툴마다 분할 기준을 정하는 데 사용하는 방법들을 다양하게 제공하고 있으며, 분할 과정을 멈추게 하기 위한 다양한 옵션을 제공하여 사용자가 지정하게 하고 있다. 의사결정나무에서 생성된 규칙은 인간이 이해하기 매우 쉬운 형태로 만들어지기 때문에 그 설명력이 매우 뛰어나다는 장점이 있다[Berry 등, 1997]. 그러나 연속형 데이터를 취급하는데 다소 어려움이 따르며, 나무 생성에 사용되는 데이터의 양에 나무의 생성 결과가 매우 민감하다[장남식 등, 1999]. 의사결정나무 기법은 전형적인 분류 및 예측 문제[Martens 등 1998; Sohn 등, 2004; He 등, 2006] 이외에도 상품 추천시스템[Cho 등, 2002], 직무태도에 대한 행위 분석[Tung 등, 2005] 등에 이용된 것에서 알 수 있듯이 광범위한 영역에서 사용되어 왔다.

3.3 사례기반 추론(Case-Based Reasoning)

인간은 현재 직면한 문제를 해결하기 위해서 과거에 해결했던 문제들 중, 현재와 가장 비슷한 상황의 문제를 기억해 낸 후, 그 문제에 사용했던 해법에 약간의 수정을 가하여 현재 직면한 문제를 해결하는 방식을 많이 사용한다고 한다[Riesbeck 등, 1989]. 사례기반 추론은 이러한 인간의 문제 해결 방식을 모방한 기계학습 방법이라고 할 수 있다. 사례기반 추론의 기본 아이디어는, 주어진 새로운 사례의 해를 구하기 위하여 우리가 답을 알고 있는 기존의 많은 사례들 중에서 새로운 사례와 유사한 사례를 골라서 그들의 답을 근간으로 새로운 사례의 해를 구하자는 것이다.

따라서, 사례기반 추론에 의하여 어떤 사례의 해를 구하기 위해서는 몇 가지 고려하여야 할 사항이 있다. 즉, 사례의 표현, 유사 사례를 구하기 위한 함수의 정의, 해를 구하기 위해 사용하려고 하는 유사 사례의 수, 유사 사례들의 해들로부터

새로운 사례의 해를 구하기 위한 함수의 정의 등을 결정하여야 한다. 먼저, 사례를 표현하는데 필요한 필드의 선택 및 형식을 결정하여야 하고, 다음으로 유사 사례를 구하기 위한 함수를 정의할 때 각 필드의 가중치를 결정하여야 한다. 또한 유사 사례들로부터 새로운 사례의 해를 구하기 위한 함수를 정의 할 때에는 유사 사례들의 가중치를 어떻게 할 것인가를 결정하여야 한다.

사례기반 추론의 경우 입력 변수나 목표 변수의 형태에 매우 자유로우며, 어느 정도 설명력이 있고, 데이터의 이상치에 그 결과가 크게 영향을 받지 않는다라는 특징이 있다. 사례기반 추론 기법은 Q&A 시스템 개발[Fu 등, 2004], 대규모 고객화 환경에서의 제품 구성[Tseng 등, 2005], 선박의 엔진실 설계[Kowalski 등, 2005], 마케팅 계획 수립[Changchien 등, 2005], 그리고 주식 시장에서의 주가 예측[Chun 등, 2005] 등 매우 다양한 분야에서 폭넓게 이용되어 왔다.

3.4 사례기반 추론과 인공신경망 및 의사결정나무 기법과의 비교

<표 3-1>은 위에서 설명한 세 가지 기법들을 몇 가지 비교 포인트에서 비교하고 있다. 인공신경망의 경우 입력 변수와 출력 변수의 형태에 있어서 기본적으로 0에서 1사이의 숫자형만 가능하다. 만약 범주형 데이터를 사용하고자 할 때에는 별도의 변환 작업을 수행해 주어야 한다. 의사결정나무의 경우에도 입력 변수에는 숫자형 데이터가 사용되어야 하며 목표 변수에는 범주형 데이터만 사용 가능하다. 변형된 의사결정나무 기법인 CART 모형은 목표형 변수에 숫자형이 사용가능 하긴 하지만 완전한 숫자형 데이터의 사용은 아니다. 그러나 사례기반 추론의 경우 입력 변수나 목표 변수의 형태에 매우 자유롭다는 특징이 있다. 따라서 숫자형과 범주형 데이터가 혼재되어 있는 데이터의 경우에는 사례기반 추론 기법을 사용하는 것이 매우 용이하다.

기법의 주요 비교 특성 중 하나로 설명력을 들 수 있다. 인공신경망의 경우 설명력이 매우 떨어지는데 반해 의사결정나무와 사례기반 추론 기법은 그 설명력이 비교적 높은 것으로 알려져 있다. 사례기반 추론 기법은 추론 과정 그 자체가 출력값에 대한 설명이 될 수 있으며, 의사결정나무의 경우도 최종적으로 생성되는 if-then 형태의 규칙은 사람이 이해하기 매우 쉬운 형태이다.

본 연구에서 사용될 데이터의 목표 변수는 숫자형 변수이면서 또 그 범위가 매우 넓은 특징을 갖고 있다. 범위가 넓을 뿐만 아니라 분포에 있어서도 매우 불규칙한 형태를 띠고 있다. 따라서 직접 과거 데이터의 값에 약간의 수정을 가하여 사용하는 사례기반 추론 기법이 우리의 실험 데이터에 가장 적합할 것으로 생각된다. 또한 입력 변수의 형태가 숫자형, 범주형, 이진형 등 다양한 형태들로 구성되어 있다. 이렇게 다양한 형태의 데이터를 모형에 사용할 때 특별한 변환 작업이 필요 없는 사례기반 추론은 매우 편리하게 이용될 수 있다.

이와 더불어 진료비 예측의 경우 환자나 병원의 입장에서 그 예측 결과를 이용하고자 할 때 해당 비용으로 예측 된 이유를 이해하는 것이 매우 중요하다. 예를 들어 병원의 입장에서 예측 모형을 병원의 정책입안이나 예산 편성 등에 이용하고자 할 때 모형의 설명력은 정책의 입안이나 예산 편성 등에 정당성을 부여할 수 있는 매우 중요한 근거가 될 수 있기 때문이다. 위와 같은 이유들로 인해 본 연구에서는 사례기반 추론 기법을 의료비 예측 모형 개발의 주요 기법으로 사용하게 되었다.

<표 3-1> 세 가지 데이터마이닝 기법의 비교

기법	입력변수의 형태	목표변수의 형태	설명력
사례기반 추론	숫자형, 범주형	숫자형, 범주형	높음
인공 신경망	기본적으로 숫자형	기본적으로 숫자형	매우 낮음
의사결정 나무	숫자형	기본적으로 숫자형(CART의 경우 범주형 가능)	매우 높음

Ⅳ. 데이터 및 실험 모형

4.1 실험 데이터

본 연구에서 사용된 데이터는 우리나라의 한 대형 대학 병원에서 수집한 암환자 데이터이다. 환자 1인당 하나씩 갖게 되는 '환자 차트'에 기록된 데이터들을 중심으로 수집하였다. 수집 기간은 2002년 10월부터 2004년 8월까지이며, 원 데이터는 총 12개의 변수로 구성되어 있다. '나이', '성별' 등 환자 개인의 신상에 관련된 데이터를 포함하여 환자의 구분은 어디에 속하는지, 어떠한 질병들을 앓고 있는지 등에 대한 기록들이 정리되어 있다. 이 이외에도 전과 횟수, 협의진단 횟수 등 환자에 대한 여러 종류의 데이터를 담고 있다. 변수의 내용은 <표 4-1>과 같다.

<표 4-1> 원 데이터 변수 내용

변수	보충 설명
나이	환자의 나이
성별	환자의 성별
환자군	일반, 보험, 교직원 등 환자가 속한 군
양성종양 개수	양성 종양의 개수
암 구분	현재 앓고 있는 암(복수가능)
기타 암 개수	주요 암 이외의 앓고 있는 기타 암의 개수
질병그룹	암 이외의 앓고 있는 질병(복수가능)
전과횟수	입원 중 다른 과로 옮긴 횟수
협의진단횟수	여러 전공의사들이 협의한 횟수
환자지불총금액	환자가 직접 부담한 금액
의료보험지불총금액	의료보험에서 지불한 금액
총금액	환자지불총금액+의료보험지불총금액

이 중에서 '암 구분', '질병그룹'은 복수의 값을 가질 수 있다. 예를 들어 '암 구분'의 경우 현재 앓고 있는 암이 '위암', '폐암', 그리고 '전립선 암'과 같이 세 개의 복수 값을 가질 수 있으며, '기타 질병 그룹'도 그 값은 WHO에서 권장하는 분류 코드[3] 값으로 저장되어 있는데, 환자가 현재 앓고 있는 암 이외의 다른 질병에 대한 코드 값이다. 따라서 여러 개의 질병을 함께 앓고 있을 때에는 여러 개 값을 가질 수 있다. 이들 변수들은 예측 모형에 입력변수로 사용되기 위하여 dummy 변수화 되어 사용될 것이다.

총 3,671건 중 이번 연구에서 사용한 데이터는 4대 암에 해당하는 데이터만 사용하였다. 전처리 이전의 암 별 데이터 건수는 위암 570건, 폐암 357건, 간암 427건 그리고 대장암 471건이다. 각 암 별 진료비의 최소값 및 최대값은 <표 4-2>와 같다. <표 4-2>에서 볼 수 있는 것과 같이 진료비 값의 범위가 매우 넓은 것을 확인할 수 있다. 폐암의 경우 최소 1만원 대에서 5억원 대에 이르는 매우 넓은 범위의 진료비를 관찰할 수 있다.

<표 4-2> 암 별 진료비 최소값, 최대값 및 평균(원)

	최소값	최대값	평균
위암	24,022	334,860,288	14,476,157.37
폐암	11,420	503,118,189	12,858,110.28
간암	30,500	507,177,296	8,937,128.79
대장암	47,344	261,886,175	13,366,297.09

4.2 실험 모형

4.2.1 데이터 전처리

원 데이터에 대한 전처리 작업은 크게 두 가지 작업이 수행되었다. 첫 번째 작업은 비정상적인 값과 극단치를 제거하는 작업이다. 비정상적인 값이라 함은 치료비가 0원 또는 마이너스 값으로 기록되어 있는 경우를 말하는데, 이런 경우는 진료 등록이나 예약 후, 환자의 여러 가지 사정으로 인해 등록이나 예약을 취소했을 때, 환불을 해 주는 규정으로 인해 발생한 데이터들이다. 따라서 이러한 데이터들은 제외 시켰다.

<표 4-2>에서 볼 수 있는 것과 같이 최소값과

3) ICD9-CM.

최대값의 범위가 수 만원에서 수 억 원으로, 매우 넓은 것을 알 수 있다. 특히, 각 암 별로 최소값과, 최대값을 이루고 있는 값들은 대부분 극단치에 속하는 것으로 판정되었다. 사례기반 추론과 인공신경망 기법을 사용한 모형에서는 극단치들을 그대로 사용할 수 있었으나 의사결정나무를 사용한 모형에서는 이들 극단치들 때문에 나무 모형이 형성될 수 없었다. 따라서 의사결정나무를 사용한 모형에서는 극단치들을 제거 한 후 의사결정나무를 생성시켰다. 이들 이상치와 극단치를 제거한 후 각 암 별 데이터 건수는 위암 569개, 폐암 350개, 간암 419개, 그리고 대장암 470개이다.

두 번째 전처리 작업은 복수개의 값을 가질 수 있는 변수들을 dummy 변수화 시켜 준 작업이다. 앞에서 언급했듯이 '암 구분', '질병그룹'은 복수개의 값을 가질 수 있다. 따라서 이 두 변수들을 각각 11개, 19개의 dummy 변수로 변환 시켜 주었다. 이와 더불어 변수 '환자군'의 경우에도 인공신경망 모형에 입력 시키기 위해 4개의 변수로 더미 변수화 시켜 주었다. 결과적으로 총 39개의 초기 입력 변수들이 생성되었고 이들 중에서 최종적으로 모형에 사용 될 변수들을 선택하기 위하여 속성 선택 작업을 수행하였다.

4.2.2 속성 선택

복수 값을 갖는 것이 가능한 속성들에 대한 dummy 변수화 이후 39개로 늘어난 속성들 중에 만약 우리의 목표 변수를 예측 하는데 도움을 주지 않는 것이 있다면 그것을 제거한 후 예측 모형을 개발 하는 것이 바람직하다. 왜냐하면 목표 변수와 큰 상관이 없음에도 불구하고 예측 모형에 포함되어 있다면 모형의 성능에 부정적인 영향을 줄 수 있기 때문이다[Dash 등, 1997].

Kohavi 등은 속성 선택의 방법을 크게 두 가지로 나누었다. 첫 번째 방법은 Wrapper 방식으로서 유도(induction) 알고리즘[4]이 임시로 선택된 속성에 대한 평가에 직접 사용되는 방법을 말한다. 두 번째 방법은 Filter 방식으로서 유도 알고리즘과는 별개로 속성 선택이 모두 끝난 후 최종적으로 선택된 속성을 유도 알고리즘에 사용하는 방법을 말한다[Kohavi 등, 1997].

본 연구에서는 아래와 같이 Relief-F 알고리즘[5]과 인공신경망 기법을 이용하여 아래와 같은 속성 선택 작업을 수행하였다.

① Relief-F 알고리즘과 인공신경망 기법에서 산출되는 각 속성의 상대적 중요도를 이용하여 목표 변수와의 관련 정도에 따라 순위를 매긴다.
② Relief-F 알고리즘에서 상대적 중요도가 0 이상으로 나타나는 속성들과 이 속성들의 개수와 동일 한 개수까지의 순위로 나타나는 인공신경망 기법에서의 속성들 중, 공통적으로 나타나는 속성들을 선택한다.

각 암 별 진료비 예측 모형을 개발하기 위해 각 모형 별로 Filter 방식을 사용한 속성 선택을 수행하였다. 그 결과 위암 데이터에서는 24개, 폐암 데이터에서는 28개, 간암 데이터에서는 19개, 그리고 대장암 데이터에서는 26개의 속성이 선택되었다. 각 암 별 속성 선택 결과는 <표 4-3>과 같다. <표 4-3>을 살펴보면 각 암 별로 선택된 속성들이 모두 똑같지는 않지만, 서로 매우 비슷한 속성 모음들로 구성된 것을 확인할 수 있다. '협의진단횟수', '기타 암 개수', 그리고 '나이'는 네 개의 암 모형 모두에서 선택되었으며 '전과횟수', '양성 종양 개수'도 세 개의 암 모형에서 중요한 변수로 선택되었다.

4) 인공신경망, 의사결정나무, 사례기반추론 등 데이터마이닝 모형에 쓰이는 본 알고리즘을 말한다.
5) 어떤 인스턴스에 있어서 자신이 속한 클래스와 자신이 속하지 않은 클래스와의 거리를 고려하여 속성의 중요도를 계산하는 방법[Kira 등, 1992].

<표 4-3> 각 암 별 선택된 속성

암	선택된 속성들
위암	질병그룹10, 전과횟수, 질병그룹17, 협의진단횟수, 질병그룹5, 질병그룹19, 질병그룹2, 질병그룹8, 질병그룹3, 질병그룹9, 환자구분4, 나이, 기타 암 개수, 질병그룹6, 질병그룹18, 양성종양 개수, 질병그룹7, 질병그룹1, 성별, 대장암 여부, 질병그룹4.
폐암	전과횟수, 협의진단횟수, 질병그룹1, 기타 암 개수, 질병그룹12, 질병그룹8, 나이, 질병그룹7, 질병그룹19, 질병그룹9, 질병그룹5, 질병그룹3, 질병그룹6, 질병그룹16, 질병그룹18, 질병그룹10, 질병그룹17.
간암	협의진단횟수, 성별, 질병그룹5, 나이, 질병그룹3, 질병그룹7, 질병그룹12, 전과횟수, 기타 암 개수, 질병그룹9, 질병그룹4, 질병그룹1, 질병그룹2, 양성종양 개수, 질병그룹8, 질병그룹18, 질병그룹17, 방광암 여부, 대장암 여부, 질병그룹16.
대장암	협의진단횟수, 전과횟수, 질병그룹8, 질병그룹16, 질병그룹19, 질병그룹17, 질병그룹18 질병그룹2, 질병그룹12, 기타 암 개수, 질병그룹6, 질병그룹7, 간암여부, 폐암여부, 나이, 질병그룹3, 질병그룹5, 질병그룹4, 질병그룹1, 질병그룹9, 질병그룹10, 위암 여부, 성별, 질병그룹15, 양성종양 개수.

4.2.3 가중치 부여

사례기반추론 기법에서 속성 선택과 함께 중요하게 여겨지는 것은 속성 가중치 부여 방법이다. 많은 속성들을 사용하여 분석할 경우 대두되는 문제는 시간적인 비용이 많이 필요하다는 것 이외에도 목표 변수와의 관련성이 서로 다른 속성들이 일정하게 같은 중요도로 사용된다면 결과에 부정적인 영향을 줄 수 있다는 것이다. 이러한 이유 때문에 관련성이 적은 속성에는 낮은 가중치를, 관련성이 많은 속성에는 높은 가중치를 적용하여 모형에 사용하고 있다. 속성 가중치 부여 방법에 대해서는 많은 방법들이 제안되어 있다[Kolodner, 1993].

속성 가중치 부여는 넓은 의미로는 속성 선택과 일맥 상통할 수 있다. 즉 가중치를 0으로 부여하면 해당 속성을 사용하지 않겠다는 것이므로 선택이 이루어지지 않은 것과 동일한 의미를 갖게 된다. 본 연구에서는 Relief-F 기법과 인공신경망 기법에서 제공해 준 상대적 중요도의 평균값을 해당 속성의 가중치로 사용하였다. 각 암 별 선택 속성들 중에서 상위 5위 이내의 가중치를 갖는 주요 속성들만 살펴 보면 <표 4-4>와 같다.

<표 4-4> 암 별 중요도 상위 5위내 속성 및 가중치

암	속성 이름	가중치
위암	질병그룹 10	0.4094
	전과횟수	0.4135
	질병그룹 17	0.2011
	협의진단횟수	0.2179
	질병그룹5	0.0430
폐암	전과횟수	0.3603
	협의진단횟수	0.3251
	질병그룹1	0.0658
	기타 암 개수	0.0423
	질병그룹12	0.0595
간암	협의진단횟수	0.4171
	성별	0.0297
	질병그룹5	0.0266
	나이	0.0545
	질병그룹3	0.0235
대장암	협의진단횟수	0.2248
	전과횟수	0.1999
	질병그룹8	0.0784
	질병그룹16	0.0526
	질병그룹19	0.0410

검색 단계에서 몇 개의 유사한 사례를 검색하느냐는 사례기반 추론 기법에 있어서 또 하나의 중요한 이슈이다. 보통 3개 또는 5개의 유사한 사례들을 검색하게 되는데, 두 모형을 모두 만들어 성능을 비교해 본 결과 3개의 사례를 선정하여 사용하는 모형이 더 좋은 성능을 보였다. 따라서

본 연구에서는 3개의 유사한 사례를 검색하여 해당 환자의 진료비 예측에 사용하였다.

4.2.4 본 연구에서 사용된 유사도 함수 및 검색된 사례 교정 방법

사례기반 추론 기법에서 사용되는 유사도에는 속성간 유사도와 사례간 유사도가 있다. 속성간 유사도는 두 비교 사례에서 각 속성들간의 유사도를 말하는 것이며, 사례간 유사도는 이들 속성간 유사도의 전체 합을 이용하여 구한 두 비교 사례간 유사도를 말한다. 본 실험에서 사용된 연속형 데이터들은 모두 0에서 1사이의 값으로 정규화 되어 사용되었다. 따라서 유클리디안 거리를 이용하여 속성별 거리를 구한 후, 이 거리 값을 1에서 빼줌으로써 각 속성간 유사도 점수를 생성하였다. 이렇게 각 속성간 유사도 점수를 구한 후, 식 (1)과 같이 이들 점수들의 총합을 이용하여 두 비교 사례간 유사도를 생성하였다. 범주형 데이터의 경우에는 완전 매칭되면 해상속성의 가중치를 유사도 점수로 부여하였고 그렇지 않으면 0점을 유사도 점수로 부여하는 방법을 사용하였다.

$C_d(a_{d1}, a_{d2}, \ldots\ldots, a_{dn})$

$C_e(a_{e1}, a_{e2}, \ldots\ldots, a_{en})$

$$Sm(C_d, C_e) = \frac{\sum_{i=1}^{n}\{(1 - |a_{di} - a_{ei}|) \times w_{fi}\}}{\sum_{i=1}^{n} w_{fi}} \quad (1)$$

C_d: 사례베이스에 있는 과거 사례
C_e: 새로운 사례
a_{di}: C_d의 i번째 속성
a_{ei}: C_e의 i번째 속성
n: 속성의 개수
$Sm(C_d, C_e)$: 사례 C_d와 C_e의 유사도
w_{fi}: i번째 속성의 속성 가중치

또한 본 연구에서는 검색 단계에서 검색된 3개의 유사한 사례들을 이용하여 교정된 제안 해를 만들 때, 두 가지 방법을 이용해서 해를 만들고 그 성능을 평가한 후 더 좋은 방법을 사용하였다. 첫 번째 방법은 식 (2)와 같이 검색된 세 개의 사례가 각각 갖고 있는 해들의 단순 평균값을 이용해서 제안될 해로 만드는 방법이며, 두 번째 방법은 식 (3)과 같이 세 개의 검색된 사례의 해에 검색될 때 사용되었던 사례별 유사도(식 (4))를 고려하여 유사도에 따른 가중치를 부여한 후, 제안될 해를 생성하는 가중 평균법을 사용하는 방법이다. 두 가지 방법을 사용하여 모형의 성능을 비교한 결과 후자의 방법이 더 좋은 것으로 판명되었다. 따라서 본 연구에서는 검색된 사례에 유사도에 따른 가중치를 부여하여 제안될 해를 생성하는 가중 평균법을 사용하였다.

$$S_p = \frac{s_1 + s_2 + s_3}{3} \quad (2)$$

$$S_p = \sum_{i=1}^{3}(s_i \times w_{ci}) \quad (3)$$

$$w_{ci} = \frac{Sm(C_i, C_e)}{\sum_{i=1}^{3} Sm(C_i, C_e)} \quad (4)$$

S_p: 제안될 해
s_i: 검색된 i 번째 사례의 해
w_{ci}: 검색된 i 번째 사례의 사례 가중치
C_i: 사례베이스에서 검색된 i 번째 사례
C_e: 새로운 사례
$Sm(C_i, C_e)$: 사례 C_i와 C_e의 유사도

본 실험에서 사용된 모형들의 목표 변수는 총 진료비를 나타내는 '총금액'으로 하였으며 모형은 각 암 별로 개발되었다. 따라서 총 4개의 사례기반 추론 모형이 만들어 졌으며, 각 암 별 데이터의 건수가 비교적 적기 때문에 각각 암 별로 5-fold Cross Validation[6]을 수행하였다.

6) 데이터 세트를 k개의 세트로 무작위로 나눈 후, 그 중 하나의 데이터 세트를 테스트 단계에서 사용하고 나

<표 4-5>와 같이 위암 진료비 예측 모형은 총 21개의 입력 변수로 구성되었으며, 데이터의 총 건수는 569건이다. 이 중 사례베이스의 사례 수는 총 건수의 80%인 455(또는 456)건이며, 테스트 사례의 개수는 총 건수의 20%인 114(또는 113)건이다. 폐암의 경우에는 총 17개의 입력 변수와 총 350건의 사례로 구성되어 있으며, 이 중 280건은 사례베이스의 사례로, 70건은 테스트 사례로 사용하였다. 간암 진료비 예측 모형은 총 20개의 입력 변수를 갖고 있으며, 사례의 총 건수는 419건 이다. 이 중 335(또는 336)건은 사례베이스의 사례로 사용하였으며 84(또는 83)건은 테스트 사례로 사용하였다. 마지막으로 대장암의 경우는 총 25개의 입력 변수와 470건의 사례 중 376건은 사례베이스의 사례로, 나머지 94건은 테스트 베이스의 사례로 사용하여 모형을 구축하였다.

<표 4-5> 각 암 별 입력 변수의 수 및 사례 건수

암	입력 변수의 수	사례베이스 사례 수	테스트 사례 수	총 사례 수
위암	21	455(또는 456)	114(또는 113)	569
폐암	17	280	70	350
간암	20	335(336)	84(또는 83)	419
대장암	25	376	94	470

사례기반 추론 모형은 MS-Visual Basic v6.0과 MS-Access 2003을 사용하여 개발하였으며, 사례기반 추론 모형의 성능과 비교하기 위한 인공신경망 모델과 CART 모형은 Clementine v8.0을 사용하였다.

V. 실험 결과 및 해석

모형의 성능은 총 두 개의 척도(MAPE: Mean Absolute Percentage Error, 피어슨 상관계수)를 사용하여 비교하였다. 이 중에서 MAPE는 실제치와 예측치 사이의 오차의 정도를 보여주는 척도이며, 피어슨 상관계수는 실제치와 예측치간의 선형성을 보고자 한 척도이다. MAPE 값은 아래 식 (5)를 이용하여 얻을 수 있다.

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\frac{Y_t - \hat{Y}_t}{Y_t} \quad (5)$$

N: 총 인스턴스의 개수
Y_t : t번째 인스턴스의 실제치
$\hat{Y}_t$: t번째 인스턴스의 예측치

<표 5-1>은 각 암 별로 5-fold cross validation을 수행한 후 MAPE 값 및 피어슨 상관계수 값의 평균을 보여주고 있다.

<표 5-1> 각 암 별 MAPE 및 피어슨 상관계수

암	모형	MAPE	피어슨 상관계수
위암	CBR	2.21	0.86
	ANN	6.07	0.80
	CART	2.27	0.82
대장암	CBR	1.56	0.82
	ANN	4.06	0.83
	CART	2.49	0.75
폐암	CBR	1.85	0.60
	ANN	4.13	0.59
	CART	4.16	0.70
간암	CBR	1.99	0.40
	ANN	2.38	0.41
	CART	2.11	0.51

MAPE 값은 네 개의 암 모형에서 모두 사례기반 추론 모형이 상대적으로 낮은 오차율을 보이고 있다. 위암 모형의 경우에는 CART 모형의 오차율보다 약간 작은 오차율 값을 보였으며 Neural Network 모형과 비교했을 때에는 다소 많은 차이의 오차율을 보이고 있다. 대장암 모형과 폐암 모형에서는 더욱 그 차이가 극명하게 드

머지 데이터 세트들은 학습단계에 사용하는 방법을 말하는데, 이 때 k개로 나누어진 각 데이터 세트들은 한 번씩 돌아가면서 테스트 세트가 된다[Kohavi, 1995].

러나고 있다. 마지막으로 간암 모형에서도 1.99로 다른 모형들의 값(2.38과 2.11) 보다 작은 오차율을 보이고 있는 것을 관찰할 수 있다.

사례기반 추론 모형의 MAPE 값 자체는 그다지 좋은 값이라 보기는 어려운 수치이다. 이렇게 높은 오차율을 보이는 이유는 모형에 사용된 입력 변수가 치료비를 충분히 반영하고 있지 못하기 때문인 것으로 추정된다. 치료비의 구성은 그 데이터의 특성상 같은 증상과 비슷한 정도의 질병이라 하더라도 환자의 경제적 상황, 치료 의지, 그리고 의사 개인의 특성 등 수 많은 다른 요인들에 의해서 다양하게 이루어 질 수 있다. 실제로 사례기반 추론 모형에서 가장 유사한 사례들이라고 가지고 온 세 개 사례의 치료비 값을 관찰해 본 결과 가장 유사한 사례와 세 번째로 유사한 사례의 치료비에 있어서도 상당한 차이가 발생하는 경우를 종종 관찰할 수 있었다. 즉, 전술한 바와 같이 우리가 확보한 데이터상에서는 비슷한 입력 변수의 값을 가지고 있는 사례임에도 불구하고 우리의 모형에 반영하지 못한 다른 요인들에 의해서 치료비의 편차가 매우 폭넓게 나타나고 있다는 것을 말해준다.

피어슨 상관계수는 위암과 대장암 모형에서 비교적 높은 상관관계를 보이고 있으나 나머지 암 모형에서는 상대적으로 작은 상관관계를 보이고 있는 것을 관찰할 수 있다. 그 이유는 폐암과 간암의 경우 그 사례의 수가 상대적으로 적기 때문이기도 하지만, 또한 세 가지 모형 모두에서 낮은 상관관계를 보이는 것으로 보아 데이터 세트의 특성에서 비롯된 것으로 추측된다.

Ⅵ. 결론 및 향후 연구 과제

다양한 의료 기록 데이터를 이용하여 네 가지 암 환자들의 진료비를 예측하는 모형을 구축하였다. 모형의 주 기법은 사례기반 추론 기법을 사용하였으며, 구축된 모형의 성능을 인공신경망 기법과 의사결정나무 기법의 성능과 비교하였다. 실험 결과 몇몇의 예외를 제외하고는 대부분의 데이터 세트에서 사례기반 추론이 실제치와 예측치 간의 오차에 있어서 인공신경망이나 의사결정나무 기법에 비해 더 좋은 성능을 보여주는 것을 확인할 수 있었다. 이처럼 사례기반 추론이 인공신경망이나 의사결정나무 기법에 비해서 더 좋은 예측 성능을 보여주는 이유는 사례기반 추론 기법의 추론 방법상의 특징 때문인 것으로 추정된다.

본 연구에서 사용된 데이터는 사례간 입력 값의 차이가 그다지 크지 않은 반면에 목표값인 비용의 값은 매우 큰 폭으로 차이가 나는 특징이 있다. 인공신경망이나 의사결정나무에서는 이렇게 작은 입력 값의 차이로는 큰 폭의 목표값을 충분히 정교하게 만들어 내지 못한 것으로 추정된다. 그러나 사례기반 추론 기법의 경우에는 검색된 사례의 목표값(비용)을 그대로 사용하면서 약간의 수정 작업만 거치기 때문에 사례베이스에 이미 담겨 있는 과거 데이터의 특성들을 그대로 반영할 수 있어 위와 같은 성질을 갖는 데이터의 경우 더 좋은 성능을 낼 수 있었던 것으로 판단된다. 따라서 데이터의 특성상, 사례별로 입력 값의 차이는 작지만 상대적으로 큰 폭의 목표값의 범위를 갖는 경우에는 사례기반 추론이 일반적으로 많이 사용되는 인공신경망 기법이나 의사결정나무 기법 보다 더 좋은 결과를 낼 수 있다는 가능성을 보여주었다.

본 연구에서 사용된 데이터는 사례기반 추론의 장점을 충분히 살릴 만큼의 다양한 형태의 데이터를 확보하고 있지 않다. 대부분의 데이터가 숫자 형 데이터였으며, dummy 변수화를 거친 이진 형 데이터들이 다수 포함되어 있었다. 따라서 사례기반 추론 기법의 장점을 좀 더 정확히 확인하기 위해서는 범주형 데이터, 문자형 데이터 등 좀 더 다양한 데이터 형태를 갖고 있는 데이터에 대한 실험이 필요하다. 또한 목표 변수로 사용된 총 비용 값의 범위가 1만원 대에서 5억원

대까지 매우 넓다. 따라서 이렇게 넓게 분포된 사례들을 좀 더 비슷한 성질을 가진 하위 그룹으로 나누고 그룹마다 예측 모형을 따로 개발해 보는 연구도 필요할 것이다.

〈참 고 문 헌〉

[1] 장남식, 홍성완, 장재호, *데이터 마이닝*, 대청 출판사, 1999.

[2] 한국 중앙 암 등록 본부, *한국중앙 암 등록 사업 연례 보고서*, 보건 복지부, 2003.

[3] Berry, M.J.A. and Linoff, G., *Data Mining Techniques*, John Wiley & Sons, Inc., 1997.

[4] Changchien, S.W. and Ming-Chin, L., "Design and Implementation of a Case-Based Reasoning System for Marketing Plans," *Expert Systems with Applications*, Vol. 28, 2005, pp. 43-53.

[5] Cho, Y.H., Jae, K.K., and Soung, H.K., "A Personalized Recommender System Based on Web Usage Mining and Decision Tree Induction," *Expert Systems with Applications*, Vol. 23, 2002, pp. 329-342.

[6] Chun, S. and Yoon-Joo, P., "Dynamic Adaptive Ensemble Case-Based Reasoning: Application to Stock Market Prediction," *Expert Systems with Application*, Vol. 28, 2005, pp. 435-443.

[7] Dash, M. and Liu, H., "Feature Selection for Classification," *Intelligent Data Analysis*, Vol. 1, 1997, pp. 131-156.

[8] DeBerard, M.S., Kevin S.M., Alan, L.C., and Edward B.H., "Presurgical Biopsychosocial Variables Predict Medical and Compensation Costs of Lumbar Fusion in Utah Workers' Compensation Patients," *The Spine Journal*, Vol. 3, 2003, pp. 420-429.

[9] Demšar, J., Zupan, B., Aoki, N., Wall, M.J., Granchi, T.H., and Beck, J.R., "Feature Mining and Predictive Model Construction from Severe Trauma Patient's Data," *International Journal of Medical Informatics*, Vol. 63, 2001, pp. 41-50.

[10] Enke, D. and Suraphan T., "The Use of Data Mining and Neural Networks for Forecasting Stock Market Returns," *Expert Systems with Applications*, Vol. 29, 2005, pp. 927-940.

[11] Fireman, B.H., Fehrenbacher, L., Gruskin E.P., and Ray, G.T., "Cost of Cancer for Patients in Cancer Clinical Trials," *Journal of the National Cancer Institute*, Vol. 92, No. 2, 2000, pp. 136-142.

[12] Fu, Y. and Ruimin, S., "GA Based CBR Approach in Q&A System," *Expert Systems with Applications*, Vol. 26, 2004, pp. 167-170.

[13] Goldman, D.P., Schoenbaum, M.L., Potosky, A.L., Weeks, J.C., Berry, S.H., Escarce, J.J., Weidmer, B.A., Kilgore, M.L., Wagle, N., Adams, J.L., Figlin, R.A., Lewis, J.H., Cohen, J., Kaplan, R., and McCabe M., "Measuring the Incremental Cost of Clinical Cancer Research," *Journal of Clinical Oncology*, Vol. 19, No. 1, 2001, pp. 105-110.

[14] He, J., Hae-Jin, H., Robert, H., Phang, C. T., and Yi, P., "Transmembrane Segments Prediction and Understanding Using Support Vector Machine and Decision Tree," *Expert Systems with Applications*, Vol. 30, 2006, pp. 64-72.

[15] Ismael M.B., Eisenstein, E.L, and Hammond, W.E., "A Comparison of Neural Network Models for the Prediction of The Cost

of Care for Acute Coronary Syndrome Patients," *Proceedings of AMIA Symposium*, 1998, pp. 533-537.

[16] Jayadevappa, R., Sumedha, C., Mark, W., Bernard, S.B., and Bruce, M., "Medical Care Cost of Patients with Prostate Cancer," *Urologic Oncology: Seminars and Original Investigations*, Vol. 23, 2005, pp. 155-162.

[17] Jeng, B., Jian-xun, C., and Ting-peng L., "Applying Data Mining to Learn System Dynamics in a Biological Model," *Expert Systems with Applications*, Vol. 30, 2006, pp. 50-58.

[18] Jenn-Lung S., Guo-Zhen W., and I-Pin C., "The Approach of Data Mining Methods for Medical Database," *Engineering in Medicine and Biology Society, Proceedings of the 23rd Annual International Conference of the IEEE*, Vol. 4, 2001, pp. 3824-3826.

[19] Jerez-Aragonés, J.M., Gómez-Ruiz, J.A., Ramos-Jiménez, G., Muñoz-Pérez, J., and Alba-Conejo, E., "A Combined Neural Network and Decision Trees Model for Prognosis of Breast Cancer Relapse," *Artificial Intelligence in Medicine*, Vol. 27, 2003, pp. 45-63.

[20] Kira, K. and Rendall, L.A., "A Practical Approach to Feature Selection," *Proceedings of the ninth International Conference on Machine Learning*, Aberdeen, Scotland, UK, San Mateo: Morgan Kaufmann, 1992, pp. 249-256.

[21] Kohavi, R., "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, In S. Wermter, E. Riloff, & G. Scheler(Eds.)," *The fourteenth international joint conference on artificial intelligence (IJCAI)*, Montreal, Quebec, Canada, San Francisco, CA: Morgan Kaufman Publishing, 1995, pp. 1137-1145.

[22] Kohavi, R. and John, G.H., "Wrappers for Feature Subset Selection," *Artificial Intelligence*, Vol. 97, 1997, pp. 273-324.

[23] Kolodner, J., *Case-Based Reasoning*, Morgan Kaufmann publishers, Inc., 1993.

[24] Kowalski, Z., Maria, M., Stefan Z., and Marcin, D., "CBR Methodology Application in an Expert System for Aided Design Ship'S Engine Room Automation," *Expert Systems with Applications*, Vol. 29, 2005, pp. 256-263.

[25] Li, W.S. and Chris, C., "SEMINT: A Tool for Identifying Attribute Correspondences in Heterogeneous Databases Using Neural Networks," *Data & Knowledge Engineering*, Vol. 33, 2000, pp. 49-84.

[26] Mangalampalli, A., Srirama, M.M., Rama, C., and Ajeet K.J., "A Neural Network Based Clinical Decision-Support System for Efficient Diagnosis and Fuzzy-Based Prescription of Gynecological Diseases Using Homoeopathic Medicinal System," *Expert Systems with Applications*, Vol. 30, 2006, pp. 109-116.

[27] Martens, J., Geert W., Jan, V., and Christophe, M., "An Initial Comparison of a Fuzzy Neural Classifier and a Decision Tree Based Classifier," *Expert Systems with Applications*, Vol. 15, 1998, pp. 375-381.

[28] Masuda, G., Sakamoto, N., and Yamamoto, R., "A Framework for Dynamic Evidence Based Medicine using Data Mining," *IEEE Symposium on Computer-Based Medical Systems*, 2002, pp. 117-122.

[29] Mitchell, and Tom, M., "Machine Learning and Data Mining," *Communications of the*

ACM, Vol. 42, No. 11, 1999.

[30] Penberthy, L., Retchin, S.M., McDonald, M.K., McClish, D.K., Desch, C.E., Riley, G.F., Smith, T.J., Hillner, B.E., and Newschaffer, C.J., "Predictors of Medicare Costs in Elderly Beneficiaries with Breast, Colorectal, Lung, or Prostate Cancer," *Health Care Management Science*, Vol. 2, 1999, pp. 146-160.

[31] Riesbeck, C.K. and Schank, R.L., *Inside Case-based Reasoning*, Lawrence Erlbaum Associates, 1989.

[32] Roche, K., Paul, N., Smuck, B., Whitehead, M., Zee, B., Pater, J., Hiatt, M.A., and Walker, H., "Factors Affecting Workload of Cancer Clinical Trials: Results of a Multicenter Study of the National Cancer Institute of Canada Clinical Trials Group," *Journal of Clinical Onclogy*, Vol. 20, No. 2, 2002, pp. 545-556.

[33] Romero, C.E. and Jiefeng, S., "Development of an Artificial Neural Network-Based Software for Prediction of Power Plant Canal Water Discharge Temperature," *Expert Systems with Applications*, Vol. 29, 2005, pp. 831-838.

[34] Sheng, W.H., Wang, J.T., Lu, D.C.T., Chie, W.C., Chen, Y.C., and Chang, S.C., "Comparative Impact of Hospital-Acquired Infections on Medical Costs, Length of Hospital Stay and Outcome between Community Hospitals and Medical Centres," *Journal of Hospital Infection*, Vol. 59, 2005, pp. 205-214.

[35] Sohn, S.Y., Yong K.J., and Seong, O.C., "Classification Models for Sequential Flight Test Results for Selecting Air Force Pilot Trainee," *Expert Systems with Applications*, Vol. 26, 2004, pp. 591-599.

[36] Tollestrup, K., Frost, F.J., Stidley, C.A., Bedrick, E., McMillan, G., Kunde, T., and Petersen, H.V., "The Excess Costs of Breast Cancer Health Care in Hispanic and Non-Hispanic Female Members of a Managed Care Organization," *Breast Cancer Research and Treatment*, Vol. 66, 2001, pp. 25-31.

[37] Tseng, H., Chien-Chen, C., and Shu-Hsuan, C., "Applying Case-Based Reasoning for Product Configuration in Mass Customization Environments," *Expert Systems with Applications*, Vol. 29, 2005, pp. 913-925.

[38] Tung, K., Ing-Chung, H., Shu-Ling, C., and Chih-Ting, S., "Mining the Generation Xers' Job Attitudes by Artificial Neural Network and Decision Tree-Empirical Evidence in Taiwan," *Expert Systems with applications*, Vol. 29, 2005, pp. 783-794.

◆ 저자소개 ◆

정석훈 (Chung, Sukhoon)
아주대학교에서 정치학사와 경영학사를 취득하고 동 대학원에서 경영정보학 석사를 취득하였다. (주)EC-miner Data Mining 팀과 (주)대우증권 IT Center에 근무하였으며, 현재 고려대학교 대학원 경영학과 박사과정에 재학 중이다. 주요 관심분야는 data mining, data warehouse 등이다.

서용무 (Suh, Youngmoo)
서울대학교 사범대학 수학과, 한국과학원 전산학과를 졸업하고, 한국과학기술 연구소 전산센터에서 연구원으로 재직 시 도미하여, University of Texas (at Austin)에서 전산학석사, 경영정보학박사를 취득한 후, 세종대학교, 건국대학교를 거쳐, 현재 고려대학교 경영대학에 재직하고 있다. 주요 관심분야는 web-based organizational computing, ontology, data warehouse, data mining 등이다.

◈ 이 논문은 2006년 1월 11일 접수하여 1차 수정을 거쳐 2006년 4월 3일 게재확정되었습니다.